Iranian Journal of Animal Science Research
Vol. 13, No. 4, Winter 2022, p. 615-627

نشریه پژوهشهای علوم دامی ایران
جلد ۱۳، شماره ۴، زمستان ۱۴۰۰، ص. ۶۲۷-۶۱۵

مقاله پژوهشی

بررسی اثر پرسپترون چند لایه در صحت انتخاب ژن های ریز RNA کرم ابریشم (*Bombyx mori*)

عاطفه سیددخت [۱*]، جواد رحمانی نیا [۲]

تاریخ دریافت: ۱۳۹۸/۱۲/۱۷
تاریخ پذیرش: ۱۳۹۹/۰۹/۲۹

سیددخت، ع.، و ج. رحمانی نیا. ۱۴۰۰. بررسی اثر پرسپترون چند لایه در صحت انتخاب ژن های ریز RNA کرم ابریشم (*Bombyx mori*) . پژوهشهای علوم دامی ایران ۱۳(۴): ۶۲۷-۶۱۵

چکیده

ریز RNA ها خانواده ای گسترده از مولکول های RNA کوتاه غیر کد کننده پروتئینی (ncRNA) و دارای وظایفی مهم در تنظیم فرآیندهای رشد در گیاهان و حیوانات هستند. مطالعات اندکی در ارتباط با ریز RNA های کرم ابریشم که از نظر اقتصادی بسیار مهم نیز هستند، با تمرکز بر شناسایی، آنالیز بیان و پیش بینی عملکرد انجام شده است. به طور کلی توالی ریز RNA ها در سرتاسر گونه ها بسیار محافظت شده هستند و از ساختار ساقه-حلقه اولیه در هسته که از ویژگی های بسیار مهم ریز RNA ها است، تولید می شوند. ریز RNA ها از مهمترین عوامل تنظیمی دخیل در سطوح پس از رونویسی پس از بیان ژن هستند که در تنظیم تعداد زیادی از فرآیندهای فیزیولوژیکی مانند رشد و نمو، متابولیسم و وقوع بیماری ها مشارکت می کنند. با اینکه هزاران ریز RNA در گونه های مختلف شناسایی شده اند، تعداد خیلی زیادی هنوز هم ناشناخته باقی مانده است. بنابراین کشف ژن های جدید ریز RNA یک گام مهم برای درک ریز RNA هایی است که مکانیسم های تنظیم پس از رونویسی را واسطه گری می کنند. روش های بیولوژیکی برای شناسایی ژن های ریز RNA ممکن است در شناسایی تشخیص ریز RNA های نادر محدودیت داشته باشند و بیشتر محدود به بافت های خاص و مراحل رشد و نموی ارگانیسم تحت آزمایش می شوند. این محدودیت ها منجر به پیشرفت روش های محاسباتی پیشرفته برای شناسایی ریز RNA های احتمالی جدید شده است. استفاده از روش های محاسباتی باعث افزایش دقت در شناسایی ریز RNA های کرم ابریشم خواهد شد. در این پژوهش، انواع مدل های محاسباتی برای شناسایی توالی های ریز RNA استفاده شد. با استفاده از داده های مناسب و استخراج ویژگی های بیولوژیکی مؤثر، عملکرد این روش ها ارزیابی شد. در مقایسه با سایر مدل های استفاده شده در این تحقیق، مدل پرسپترون چند لایه با بیشترین مقادیر دقت، معیار F و ضریب همبستگی متیو به عنوان روشی مناسب جهت پیش بینی توالی های ریز RNA در کرم ابریشم معرفی شد.

واژه های کلیدی: روش های محاسباتی، عوامل تنظیمی، کرم ابریشم، ریز RNA

مقدمه

تلاش های اخیر برای رهگیری ژن های ریز RNA سبب کشف هزاران ریز RNA در گونه های مختلف شده است، اما بسیاری از آن ها هنوز شناسایی نشده اند (۶). بیشتر این کوشش ها از روش های آزمایشگاهی مانند همسانه سازی مستقیم RNA های کوچک با منشأ داخلی که زمان بر، پر هزینه و نیازمند کار بسیار هستند، استفاده می کنند (۳۳). این واقعیت که ریز RNA ها در انواع سلول های ویژه در سطوح پایین یا فقط در یک شرایط خاص بیان می شوند و تشخیص

۱-بخش تحقیقات علوم دامی، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان خراسان رضوی، سازمان تحقیقات، آموزش و ترویج کشاورزی، مشهد، ایران.
۲-مؤسسه تحقیقات علوم دامی کشور، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران.

(*- نویسنده مسئول: Email: a.seyeddokht@areeo.ac.ir)
Doi:10.22067/ijasr.2020.38276.0

آزمایشگاهی آن ها را پیچیده می کند، نشان دهنده ناکافی بودن روش های آزمایشگاهی جهت شناسایی ریز RNA های جدید است. برای رفع این مشکلات چندین روش محاسباتی جهت تشخیص ژن های ریز RNA طراحی و استفاده شده است.

شناسایی ریز RNA ها پیچیده است و به یک استراتژی بین رشته ای نیاز دارد. پیشرفت های کنونی فن آوری مانند تعیین توالی پربازده سبب شده است تا الگوهای بیان آن ها آسان تر شناسایی شود (۲۹). در سال های اخیر رویکردهای بیوانفورماتیکی و زیست شناختی امکان کشف هزاران ریز RNA را در گیاهان (۴۹)، حیوانات (۱)، یوکاریوت های تک سلولی (۲۶، ۴۲) و ویروس ها (۲۲، ۴۵) فراهم کرده اند. بسیاری از توالی های آزمایشگاهی آن ها، اکنون در miRBase، مخزن اصلی توالی ها و حاشیه نویسی ریز RNA ها جمع آوری شده اند. آخرین نسخه miRBase نسخه ۲۲ می باشد که شامل تعداد ۳۵۵۸۹ ریز RNA پیش ساز مربوط به ۲۷۱ گونه می باشد (۱۸). راهکارهای معمولی که برای کشف ریز RNA ها استفاده می شود، شامل همسانه سازی (۶)، وسترن بلات (۴۶)، ریزآرایه (۲۵) و هیبریداسیون درجا (۳۲) است که زمان بر و پر هزینه هستند (۲۸). فن آوری تعیین توالی نسل بعدی (NGS)، روشی قابل اعتماد و حساس برای تعیین مقدار ریز RNA های شناخته شده و شناسایی مواردی است که کمتر رایج هستند (۴). بسیاری از الگوریتم ها برای کشف ریز RNA های جدید از داده های NGS استفاده می کنند. الگوریتم های محاسباتی برای افزایش دقت در روش های آزمایشگاهی جهت شناسایی و اعتبارسنجی ریز RNA های جدید سازگار شده اند. این ابزارها برخی از ویژگی های مهم ریز RNA را مانند حفاظت توالی در بین گونه ها و ویژگی های ساختاری مانند ساختار سنجاق سر و حداقل انرژی آزاد فولدینگ را در نظر می گیرند (۲۱). چندین ابزار مانند RNAfold و Mfold برای به دست آوردن ساختار ثانویه مبتنی بر حداقل انرژی آزاد، ابزارهای محاسباتی اصلی هستند که برای یافتن ژن های ریز RNA مورد استفاده قرار می گیرند. روش های بیوانفورماتیک مقدماتی، ریز RNA های مورد نظر در توالی های ژنوم را با هدف قرار دادن ساختار RNA ثانویه، یعنی ساختارهای سنجاق سر محافظت شده که مشخصه توالی های پیش ساز ریز RNA در گونه های مرتبط هستند، پیش بینی می کنند. MiRscan (۲۳) و miRseeker (۱۹) اصلی ترین ابزارهایی هستند که توالی های درون ژنی محافظت شده را که می توانند ساختارهای سنجاق سر را براساس RNAfold و Mfold تشکیل دهند، مورد هدف و شناسایی قرار می دهند. آن گاه MiRscan ساختارهای شناسایی شده را با ویژگی های ریز RNA های شناخته شده همانند حفاظت ساقه '۳ و '۵ مقایسه می کند، در حالی که miRseeker سنجاق سرهایی را که الگوهای واگرایی نوکلئوتیدی مشابه را با مجموعه مرجع به اشتراک می گذارند، انتخاب می کند. MiRscan و miRseeker اولین مرتبه برای شناسایی ریز RNA ها در نماتدها و مگس ها استفاده شدند و تعداد زیادی از کاندیدهای پیش بینی شده به صورت آزمایشگاهی تأیید شدند. با این حال، از آن جایی که ابزارهای مبتنی بر روش های مقایسه ای اساساً روی ریز RNA های محافظت شده تکامل یافته تمرکز می کنند، محدود به کشف ریز RNA های جدید هستند. در ادامه، راهکارهای یادگیری ماشینی برای پیش بینی ریز RNA های جدید ابداع شده است. این راهکارها با گسترش تجزیه و تحلیل های فراتر از خصوصیات توالی و ساختار، پیش بینی ریز RNA های ناشناخته را بهبود بخشیده اند. الگوریتم های فراگیری ماشینی به برنامه های رایانه ای اجازه می دهند تا از اطلاعات جمع آوری شده از ریز RNA های تأیید شده قبلی به عنوان استانداردهای ریز RNA مثبت استفاده کنند. الگوریتم هایی همچون مدل پنهان مارکوف (HMM)، طبقه بندی کننده بیز (NBC) و ماشین بردار پشتیبان (SVM) (۲۸) از این دسته اند. روش های HMM شناخت الگو را در بین مجموعه داده ها به ویژه توالی های نوکلئوتیدی ارائه می دهند (۲). مدل NBC، مدلی طبقه بندی کننده است که با استفاده از روشی نسبتاً ساده در یک مجموعه داده آموزشی بدست آمده است. NBC احتمال اینکه محاسبه یک نمونه متعلق به طبقه ای خاصی باشد (۵۱) را محاسبه می کند. SVM طبقه بندی کننده ای است که اشیاء را بر اساس مجموعه ای از ویژگی ها برای هر شی دسته بندی می کند. این طبقه بندی کننده، بردارها را از یک طبقه مثبت و یک طبقه منفی مقایسه می کند و یک ابر صفحه[1] برای تولید بهترین حاشیه جدا کننده بین آن ها فراهم می کند (۵،۴۰). ابزارهای مختلفی براساس این رویکردها برای پیش بینی ریز RNA ها از گونه های مختلف تهیه شده است. به عنوان مثال، HMM مبتنی بر ابزار ProMir (۳۱) یک مدل احتمالاتی یادگیری مشترک مبتنی بر توالی های محافظت شده و ساختارهای ثانویه است که برای پیش بینی ژن های ریز RNA انسانی استفاده می شود. نسخه بهبود یافته ProMiR II (۳۰) ، معیارهای فیلتر اضافی دیگری مانند نسبت G/C، نمره حفاظت، آنتروپی و انرژی آزاد توالی های کاندید را ارائه می دهد. پیش بینی ژن های ریز RNA محافظت شده و غیر محافظت شده همچنین با تنظیم معیارهای فیلتر امکان پذیر است. استفاده از مجموعه های داده های آموزشی مناسب، امکان استفاده از همه گونه ها را فراهم می کند. MiRRim (۴۳) یکی دیگر از ابزارهای مبتنی بر HMM است که هر دو ویژگی ساختار تکاملی و ثانویه ژن ریز RNA را برای دستیابی به توالی های جدید در نظر می گیرد. این ابزار شناسایی با کارآیی زیادی ریز RNA های جدید انسانی، به ویژه آن هایی که با ریز RNA های شناخته شده خوشه بندی شده اند، را در نظر می گیرد. HHMMiR (۱۷) به صورت *de-*

1- Hyperplane

novo سنجاق سرهای ریز RNA را در غیاب حفاظت تکاملی پیش بینی می کند. در این روش از یک HMM سلسله مراتبی بهره گیری می شود که از توالی مبتنی بر ناحیه به عنوان اطلاعات ساختاری پیش سازهای ریز RNA استفاده می کند. ابزار دیگری که آزادانه در دسترس است، ابزار پیش بینی SSCprofiler (۳۵) می باشد که از یک روش احتمالی مبتنی بر HMM برای پروفایل آموزش داده شده، در تشخیص ریز RNA ها استفاده می کند.

جمع آوری مجموعه داده های مناسب برای اغلب الگوریتم های یادگیری ماشین برای تولید یک طبقه بندی گر خوب آموزش دیده، بسیار ضروری است. اگر توالی ها خیلی مصنوعی باشند، آن گاه احتمال زیادی وجود دارد که روش یادگیری ماشین برای اینکه بین ریز RNA های واقعی و توالی های غیر ریز RNA تفکیک قائل شود، به اندازه کافی آموزش داده نشود (۴۸). از سویی دیگر، اگر مجموعه داده های منفی دارای تشابه زیادی با مجموعه داده های مثبت باشند، روش یادگیری ماشین ناتوان از تمایز بین این دو مجموعه داده خواهد بود (۴۸).

در پژوهش حاضر، مدل های طبقه بندی کننده مختلفی پیشنهاد شده اند که قادر به طبقه بندی ریز RNA های پیش ساز واقعی از ساختارهای سنجاق سر کاذب[1] و همچنین از سایر RNA های غیر کد کننده می باشند.

مواد و روش‌ها

داده های مثبت برای پیش بینی ژن های ریز RNA از نسخه ۲۲ miRBase(۱۸) استخراج شد. با این حال برخی از ورودی ها در سایر پایگاه های داده وجود دارد که به عنوان ریز RNA ها پیشنهاد می شوند، اما آن ها خواص لازم برای طبقه بندی به عنوان ریز RNA ها را از جمله داشتن بیش از یک حلقه برآورده نمی کنند. مشخص شده است که مجموعه توالی هایی که به عنوان کنترل مثبت های مرجع، از miRBase گرفته شده است، برای ایجاد یک مجموعه با اطمینان بالا برای استفاده به عنوان کنترل مثبت دارای اعتبار می باشد. همچنین مشخص شده است که دقت پیش بینی پس از فیلتر ریز RNA های غیر محتمل می تواند بهبود یابد (۳۹). به جز این مشکلات کوچک، در مطالعات پیش بینی ژن ریز RNA، انتخاب نمونه های مثبت (به عنوان نمونه با استفاده از نمونه های ریز RNA های شناخته شده) معمولاً مشکل نیست، در حالی که ایجاد نمونه های منفی چالش برانگیز است (۲۴). در پژوهش حاضر داده های مثبت شامل ۳۹۰ توالی های ریز RNA پیش ساز کرم ابریشم (*Bombyx mori*) است که از پایگاه miRBase نسخه ۲۲ (۱۸) استخراج شده است.

ویژگی های استفاده شده، شامل ویژگی های ترمودینامیکی مانند حداقل انرژی آزاد با شاخص های مختلف، ویژگی های موقعیتی و درصد G+C می باشد. کد های لازم برای محاسبه این ویژگی ها در نرم افزار C# پیاده سازی شد. ویژگی های ترمودینامیکی و ساختار ثانویه با استفاده از وب سرور RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) استخراج شد.

همچنین ساختار ثانویه، حداقل انرژی آزاد (MFE) و ویژگی های وابسته به جفت بازی شامل تعداد جفت بازها و میانگین جفت بازها به ازای هر توالی، محاسبه و برای آموزش مدل ها مورد استفاده قرار گرفت.

در ساختار ثانویه پیش بینی شده، برای هر نوکلئوتید فقط دو موقعیت جفت شده یا جفت نشده وجود دارد و به ترتیب توسط براکت های ')' یا '(' و نقطه '. ' نشان داده می شود. براکت چپ ')' نشان می دهد که نوکلئوتید جفت شده در نزدیک انتهای ´۵ قرار دارد و با یک نوکلئوتید دیگر در انتهای ´۳ می تواند جفت شود که به عنوان یک براکت راست '(' نشان داده می شود. در RNAFold این دو وضعیت از هم تفکیک نشده است و از ')' برای هر دو موقعیت استفاده شده است. برای هر سه نوکلئوتید مجاور، هشت ترکیب ساختار احتمالی ($۲^۳$) وجود دارد: ')))' ، '.))'، '..)'، '...'، ')).'، ')..'، '.).'و ').)'. با در نظر گرفتن یک نوکلئوتید میانی در میان این سه نوکلئوتید، (۸×۴) ۳۲ ترکیب از توالی های ساختاری وجود دارد، که به شکل ')))U'، 'A)).' و غیره علامت گذاری می شود (۱۶) (شکل ۱).

معیار های ویژگی های مورد استفاده

یکی از معیارهایی که توالی های RNA کوتاه باید دارا باشند تا بتوانند به عنوان ریز RNA طبقه بندی شوند، این است که باید توسط آنزیم Dicer شناخته و پردازش شده باشد. این در حالی است که وقتی یک مجموعه داده منفی تعریف می شود، این معیار باید برای اینکه کنترل منفی های انتخاب شده توسط آنزیم Dicer تشخیص داده نشوند، به طور مؤثر مورد استفاده قرار گیرد (۳۷). بنابراین این فرض که سنجاق سرهای مناطق اگزونی داده های منفی خوبی هستند، بسیار غیر قابل اطمینان است. در این پژوهش از چند نوع توالی های RNA غیر کد کننده کوتاه، به عنوان توالی های منفی استفاده شد.

مدل های محاسباتی

مدل های آموزشی مختلفی با استفاده از ویژگی های محاسبه شده برای توالی های ریز RNA کرم ابریشم توسط نرم افزار وکا مورد تجزیه و تحلیل قرار گرفتند که عبارتند از:

1- pseudo hairpins

شبکه بیزی (Bayesian network)

یک گراف جهت دار غیرمدور است که مجموعه‌ای از متغیرهای تصادفی و نحوه ارتباط مستقل آن‌ها را نشان می‌دهد. شبکه بیزین یک ابزار مناسب برای شناسایی روابط احتمالی به منظور پیشگویی یا ارزیابی کلاس عضویت است (۲۷).

دسته‌بندی کننده بیز ساده (Naive Bayes classifier)

در یادگیری ماشین به گروهی از دسته‌بندی‌کننده‌های ساده بر پایه احتمالات گفته می‌شود که با فرض استقلال متغیرهای تصادفی و براساس قضیه بیز ساخته می‌شوند. به طور ساده روش بیز روشی برای دسته‌بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده است. این روش از ساده ترین الگوریتم های پیش‌بینی است که دقت قابل قبولی هم دارد (۱۵).

پرسپترون چند لایه (Multilayer perceptron)

دسته ای از شبکه‌های عصبی مصنوعی پیشخور است. یک MLP شامل حداقل سه لایه گره است: یک لایه ورودی، یک لایه پنهان و یک لایه خروجی. به جز گره‌های ورودی، هر گره یک نورون است که از یک تابع فعال‌سازی غیر خطی استفاده می‌کند.MLP از تکنیک یادگیری نظارت شده به نام باز پرداخت برای آموزش استفاده می‌کند. لایه‌های متعدد آن و فعال‌سازی غیر خطی آن MLP را از یک پرسپترون خطی متمایز می‌کند. در واقع می تواند داده هایی را متمایز کند که به صورت خطی قابل تفکیک نیستند (۴۱).

گرادیان کاهشی تصادفی ((Stochastic Gradient (SGD)

روشی مبتنی بر تکرار برای بهینه سازی یک تابع مشتق پذیر به نام تابع هدف (تابع هزینه) است که یک تقریب تصادفی از روش گرادیان کاهشی می‌باشد. در حقیقت گرادیان کاهشی تصادفی الگوریتمی در اختیار ما قرار می‌دهد که طی چند حلقه تکرار مقدار کمینه یک تابع و مقادیری را که به ازای آنها تابع کمینه مقدار خود را می گیرد، بدست بیاوریم (۵۳).

رگرسیون لجستیک (Logistic regression)

یک مدل آماری رگرسیون برای متغیرهای وابسته دوسویی است. این مدل را می‌توان به عنوان مدل خطی تعمیم یافته‌ای که از تابع لوجیت به عنوان تابع پیوند استفاده می‌کند و خطایش از توزیع چند جمله‌ای پیروی می‌کند، به‌حساب آورد. منظور از دو سویی بودن، رخ داد یک واقعه تصادفی در دو موقعیت ممکنه است. زمانی که متغیر وابسته ما دو وجهی است و می‌خواهیم از طریق ترکیبی از توابع منطقی دست به پیش بینی بزنیم، از رگرسیون لجستیک استفاده می کنیم (۳۴).

یادگیری درخت تصمیم (Decision tree learning)

گروهی از الگوریتم های یادگیری ماشین هستند که در طبقه‌بندی آماری کاربرد دارند. درخت‌های تصمیم به گروه الگوریتم های یادگیری تحت نظارت تعلق دارند و بیشتر آنها بر اساس حداقل سازی کمیتی به نام آنتروپی ساخته می‌شوند. هرچند توابع دیگری هم برای یادگیری درخت تصمیم وجود دارند. نمونه های قدیمی درخت تصمیم تنها قادر به استفاده از متغیرهای گسسته بودند، اما الگوریتم های جدیدتر هر دو نوع متغیر گسسته و پیوسته را در یادگیری به کار می‌برند. یکی از مزایای مهم الگوریتم درخت تصمیم قابلیت فهم و تفسیر آسان است که محبوبیت این الگوریتم را بالا برده است (۳).

رگرسیون (Regression)

رگرسیون یک نوع مدل آماری است که برای پیش‌بینی یک متغیر از روی یک یا چند متغیر دیگر استفاده می شود. رگرسیون خطی نوعی تابع پیش بینی کننده خطی است که در آن متغیر وابسته، متغیری که قرار است پیش بینی شود، به صورت ترکیبی خطی از متغیرهای مستقل پیش بینی می شود، بدین معنی که هر کدام از متغیرهای مستقل در ضریبی که در فرایند تخمین برای آن متغیر به دست آمده ضرب می شود. جواب نهایی مجموع حاصل‌ضرب ها به علاوه یک مقدار ثابت خواهد بود که آن هم در فرایند تخمین به دست آمده است (۵۰).

آدابوست (AdaBoost)

این الگوریتم یک متا الگوریتم تطبیقی بوده که به منظور ارتقاء عملکرد و رفع مشکل رده های نامتوازن، همراه دیگر الگوریتم های یادگیری استفاده می‌شود. در این الگوریتم، طبقه بند هر مرحله جدید به نفع نمونه های غلط طبقه بندی شده در مراحل قبل تنظیم می‌گردد. آدابوست نسبت به داده‌های نویزی و پرت حساس است. ولی نسبت به مشکل بیش برازش از بیشتر الگوریتم های یادگیری برتری دارد. طبقه بند پایه که در اینجا استفاده می‌شود، فقط کافیست از طبقه‌بند تصادفی (۵۰ درصد) بهتر باشد و به این ترتیب بهبود عملکرد الگوریتم با تکرارهای بیشتر بهبود می یابد. حتی طبقه بندهای با خطای بالاتر از تصادفی با گرفتن ضریب منفی عملکرد کلی را بهبود می بخشند (۵۷).

اعتبارسنجی داده های مورد آنالیز

در این پژوهش از روش اعتبار سنجی متقاطع ده تایی برای تأیید داده های مورد آنالیز، استفاده شد. در این نوع اعتبارسنجی داده‌ها به K زیر مجموعه افراز می‌شوند. از این K زیر مجموعه، هر بار یکی برای آزمون و K-1 مورد دیگر برای آموزش به کار می‌روند. این روال K مرتبه تکرار می‌شود و همه داده‌ها دقیقاً یک بار برای آموزش و یک بار

برای آزمون به کار می‌روند. در نهایت میانگین نتیجه این K بار اعتبارسنجی به عنوان یک تخمین نهایی برگزیده می‌شود.

نتایج

بررسی ویژگی های ساختاری و ترمودینامیکی برای توالی های ریز RNA

خروجی نرم افزار RNAfold به عنوان نمونه، برای ریز RNA مربوط به let-7 (bmo-let-7 MI0004968) در شکل ۱ آورده شده است.

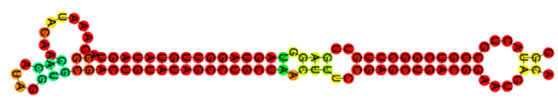
از این خروجی ها به عنوان ورودی به ازای هر توالی برای آموزش مدل های آموزشی استفاده شد.

```
1
GGUACUGCCGUCGGCUUGUUGAGGUAGUAGGUUGUAUAGUACGGAAAUACAACACAUAGGUGCGACUGUAUAGCCUGCUA
ACUUUCCGAGCUGACGGAAUGACA
1
.((....(((((((((((..(((.((((((((((((((((.((...................))))))))))))))))))
.)))..)))))))))))....)).
```

شکل ۱– ساختار ثانویه برای توالی ریز RNA جایگاه Let-7 کرم ابریشم در نماد نقطه براکت (dot-bracket) با حداقل انرژی آزاد kcal / mol ۴۶/۳۰–

Figure 1- Secondary structure for Let-7 microRNA sequence of silkworm in dot-bracket symbols with a minimum free energy of -46/30 kcal/mol

شکل ۲، ساختار ثانویه برای توالی ریز RNA جایگاه Let-7 کرم ابریشم را نشان می دهد. رنگ ها در شکل ۲ نشان دهنده اندازه گیری های اعتبارسنجی برای موقعیت هر باز می باشد و هر باز توسط آنتروپی موقعیتی رنگ آمیزی شده است.



شکل ۲– ساختار ثانویه برای توالی ریز RNA جایگاه Let-7 کرم ابریشم

Figure 2- Secondary structure for Let-7 silkworm microRNA sequence

مدل های یادگیری ماشین

یادگیری ماشین در بسیاری از برنامه ها و مطالعات بیوانفورماتیک استفاده می شود. ایجاد روزافزون داده هایی که توسط تکنیک های زیست شناسی مولکولی مدرن ایجاد شده است، باعث شده تا نیاز به الگوریتم های طبقه بندی و پیش بینی دقیق احساس شود و از زمان دستیابی به آن ها روش های سنتی دیگر عملی نیست (۷). تعداد بیشماری از رشته های زیست شناسی از روش های یادگیری ماشین برای استخراج اطلاعات از داده هایی مانند ژنومیک، سیستم بیولوژی، تکامل، ریزآرای و پروتئومیکس (۲۰) استفاده می کنند. مدل های یادگیری ماشین در تشخیص ریز RNA ها با استفاده از الگوریتم های پیش بینی متمایز، مبتنی بر استفاده از قوانین متفاوتی می باشد. قوانینی مانند تعیین اینکه آیا یک توالی مشخص، ریز RNA است و به صورت دستی ایجاد نشده اند. در نتیجه این قوانین متناسب و آموزش دیده هستند و یا از نمونه ها آموزش دیده اند (۲۰). معمولاً روش های مبتنی بر یادگیری ماشینی با فرایند یادگیری توالی، ساختار یا ویژگی های ترمودینامیکی ریز RNA ها شروع می شوند. در مرحله بعد یک طبقه بندی کننده مبتنی بر اطلاعات بدست آمده از طریق مجموعه داده های مثبت و منفی برای تصمیم گیری در مورد اینکه توالی های ناشناخته، ریز RNA های واقعی هستند، تشکیل می شود و پارامترهای حاصل، مجموعه ای از ویژگی های عددی هستند که ریز RNA های کاندید را براساس به عنوان مثال حداقل انرژی آزاد پیچش آن ها تعریف می کنند.

یادگیری ماشین برای پیش بینی ژن ریز RNA به طور انحصاری، مبتنی بر یادگیری با نظارت است که در آن یک الگوریتم برای یادگیری آموزش داده می شود، یعنی یک تابع که داده های ورودی را به داده های خروجی مورد نیاز نگاشت می کند (۵۴). معمولاً مجموعه ای از پارامترهای تعیین کننده یک کاندید (به عنوان مثال، MFE، تعداد دی نوکلئوتیدها، طول ساقه و غیره) به عنوان ورودی و ریز RNA یا غیر ریز RNA خروجی هستند. با وجود این که خروجی پیش بینی شده ناشناخته است، این دستگاه توسط ورودی آموزش داده می شود. ایده اصلی این فرایند این است که یادگیرنده ماشین باید توانایی ساده کردن

این نمونه ها را داشته باشد (داده های ورودی : نمونه های مثبت و منفی) و به درستی کاندیدها را طبقه بندی کند (۲۴). مهمترین عاملی که بر صحت نتایج تأثیر می گذارد، انتخاب ویژگی ها است، زیرا پارامتر کردن نمونه ها به ویژگی ها به طور خودکار (۱۰، ۲۴) انجام نمی شود. برای آزمون صحت و دقت فرآیند یادگیری ماشین، یک سیستم موسوم به اعتبار سنجی متقاطع استفاده شد. اعتبار سنجی متقاطع برای جلوگیری از خطاهای نوع III –رد اشتباه فرض صفر– مهم است (۳۸). اعتبار سنجی متقاطع شامل تقسیم یک نمونه از داده ها به زیر مجموعه های مربوطه و انجام تجزیه و تحلیل در یک زیر مجموعه (مجموعه آموزش یا یادگیری) و اعتبار یابی تجزیه و تحلیل بر مجموعه مورد آزمون است. مجموعه های نمونه را می توان به درصدهای مشخص (به عنوان مثال ۷۰ درصد از نمونه های موجود در مجموعه یادگیری، ۳۰ درصد باقی مانده در مجموعه تست) تقسیم کرد. نکته اساسی این است که این مجموعه داده ها نباید دارای نمونه های مشترک باشند.

پس از اعتبارسنجی متقاطع، بهترین مدل برای انجام پیش بینی ها انتخاب و استفاده خواهد شد.

خروجی مدل ها

نتایج خروجی الگوریتم های بیز، لجستیک، پرسپترون چند لایه، گرادیان کاهشی تصادفی، ریشه های تصمیم، رگرسیون و آدابوست در جدول ۱ آمده است.

با توجه به مقادیر برآوردهای روش‌های مختلف در شناسایی ریز RNA ها می‌توان نتیجه گرفت که بین روش‌های لجستیک، بیزی، آدابوست، پرسپترون چند لایه، درخت تصمیم گیری، گرادیان کاهشی تصادفی و رگرسیون، مدل پرسپترون چند لایه با بیشترین میزان MCC و کمترین مقدار مثبت کاذب (شکل ۴) به عنوان یک روش برتر در شناسایی ریز RNA های کرم ابریشم نسبت به سایر روش‌ها انتخاب می شود.

جدول۱– آنالیز مدل های مختلف برای ریز RNA های پیش ساز کرم ابریشم

Table 1- Analysis of different models for silkworm precursor microRNAs

روش ها Methods	دقت Accuracy	معیار F F-Measure	ضریب همبستگی متیو Matthews Correlation Coefficient	ناحیه ROC[1] ROC Area	ناحیه PRC[2] PRC Area
شبکه بیز Bayes Net	۰/۸۶۴	۰/۸۶۳	۰/۷۲۵	۰/۹۴۸	۰/۹۵۰
بیز ساده Naive Bayes	۰/۷۸۴	۰/۷۳۵	۰/۵۲۷	۰/۸۳۷	۰/۸۲۸
لجستیک Logistic	۰/۹۱۱	۰/۹۰۹	۰/۸۲۰	۰/۹۵۶	۰/۹۴۸
پرسپترون چند لایه Multi layer Perceptron	۰/۹۶۲	۰/۹۶۱	۰/۹۲۳	۰/۹۷۵	۰/۹۶۶
گرادیان کاهشی تصادفی Stochastic gradient descent	۰/۸۸۹	۰/۸۸۴	۰/۷۷۳	۰/۸۸۵	۰/۸۴۲
لجستیک ساده Simple Logistic	۰/۹۱۲	۰/۹۱۰	۰/۸۲۲	۰/۹۵۱	۰/۹۴۵
ریشه تصمیم Decision Stump	۰/۷۳۴	۰/۷۱۶	۰/۴۵۴	۰/۷۲۰	۰/۶۶۲
رگرسیون Regression	۰/۹۲۵	۰/۹۲۴	۰/۸۴۹	۰/۹۸۴	۰/۹۸۵
آدابوست Ada Boost	۰/۸۳۳	۰/۸۳۲	۰/۶۶۵	۰/۹۰۸	۰/۹۰۰

۱. ناحیه منحنی مشخصه عملکرد سیستم یا منحنی عملیاتی گیرنده.

۲. ناحیه منحنی دقت

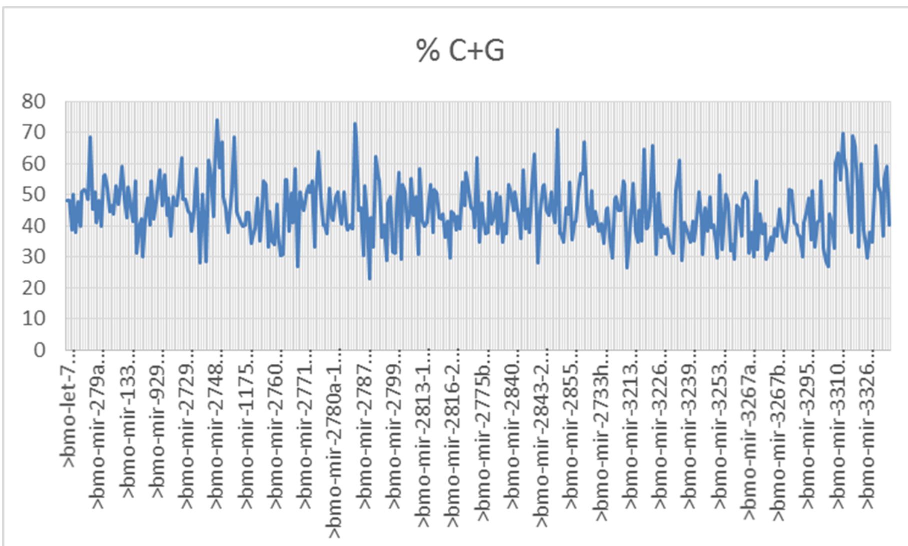
1. Receiver Operating Characteristic curve.
2. Precision-Recall curve.

با توجه به جدول ۱، بهترین مدل پیش بینی کننده، مدل پرسپترون چند لایه با دقت ۰/۹۶۲ و ضریب همبستگی ۰/۹۲۳ تعیین شد. پارامترهای این مدل شامل نرخ یادگیری ۰/۳، نرخ حرکت ۰/۲، زمان آموزش ۵۰۰ و آستانه تأیید ۲۰ در نظر گرفته شدند. تابع نرون های لایه خروجی این مدل برخلاف تابع خطی شبکه پیشرو پس انتشار، تابع tansig می‌باشد. به علاوه تعداد نرون های لایه خروجی برابر تعداد

کلاس ها درنظر گرفته شده و ویژگی ها به صورت برداری برای آموزش به شبکه معرفی می شوند. در طی این آنالیز از داده های اعتبار سنجی برای ارزیابی میزان بهینه بودن ساختار یک شبکه نسبت به شبکه‌ای دیگر استفاده می‌شود. مقادیر اولیه پارامترهای شبکه نقش مهمی در حرکت شبکه به سمت جواب بهینه دارند. با توجه به اینکه داده های آموزشی ممکن است نتیجه مدل را اریب کنند، از روش اعتبارسنجی ضربدری۱۰ تایی استفاده شد.

اختلاف نظر بسیاری در مورد اهمیت محتوای نوکلئوتیدی در پردازش ریز RNA وجود دارد. در این پژوهش با توجه با نتایج می توان بیان نمود که محتوای نوکلئوتیدهای G و C پارامترهای اصلی برای شناسایی توالی های ریز RNA های پیش ساز می باشند (شکل ۳) و احتمالاً به این دلیل است که مدل های آموزشی به جای آنالیز این پارامترها به طور مستقل، آن ها را در فرایند آموزشی ترکیب می کنند.

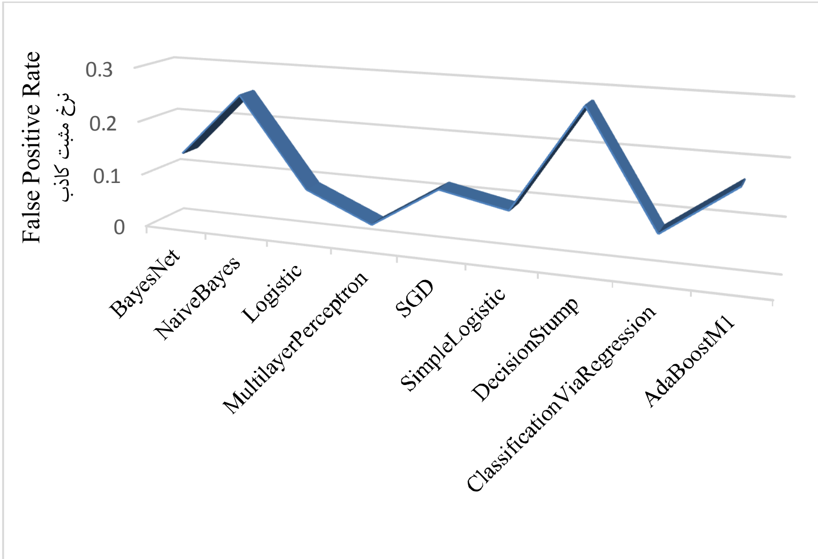
شکل ۳ محتوای نوکلئوتیدی ریز RNA های کرم ابریشم را از نظر درصد گوانین و سیتوزین نشان می دهد. این ویژگی در ردیابی مولکول های ریز RNA به عنوان یک خصوصیت مؤثر به حساب می آید. این شکل به ویژگی های شناسایی شده در توالی های ریز RNA کرم ابریشم اشاره می کند و نشان می دهد درصد نوکلئوتیدهای G+C یک ویژگی مهم در شناسایی ریز RNA ها است.



شکل۳- درصد نوکلئوتیدهای G+C در تعدادی از توالی های ریز RNA کرم ابریشم

Figure 3- Percentage of G + C nucleotides in some of silkworm microRNA sequences

همچنین با توجه به برآورد مقادیر نرخ مثبت کاذب که در شکل ۴ آورده شده است، می توان نتیجه گرفت که مدل پرسپترون چند لایه با نرخ مثبت کاذب ۰/۰۳۸ دارای کمترین مقدار مثبت کاذب و درنتیجه دارای بیشترین میزان دقت می باشد.



شکل ٤- مقایسه نرخ مثبت کاذب در انواع الگوریتم های آموزشی

Figure 4- Comparison of false positive rates in a variety of training algorithms

بحث

در این پژوهش مدل ها و ویژگی های ساختاری مختلف برای پیش بینی ریز RNA ها استفاده شد. نتایج مطالعه ما نشان داد که مدل پرسپترون چند لایه این توانایی را دارد که ویژگی ها را از توالی های RNA یاد بگیرد و در نتیجه با دقت بالا برای تشخیص محاسباتی ریز RNA های کرم ابریشم استفاده شود. در آزمایش ما، از توالی های RNA های غیر کد کننده برای انجام پیش بینی استفاده شد. انتخاب داده های مناسب برای عملکرد بهینه مدل ها ضروری است، اگرچه مدل های یادگیری می توانند ویژگی ها را از داده ها یاد بگیرند. به دلیل کوچک بودن مجموعه داده، کیفیت داده ها و روش بردارسازی از توالی های ورودی تأثیر زیادی در عملکرد طبقه بندی کننده دارد. نتایج نشان داد که مدل پرسپترون چند لایه پیشنهاد داده شده در این پژوهش می تواند با موفقیت بیشتری نسبت به سایر مدل ها در مجموعه داده های آموزشی، استفاده شود. نرخ خطای پیش بینی کم در مجموعه داده های این پژوهش، نشان دهنده توانایی تعمیم بالای مدل پیش بینی مورد استفاده است.

پژوهش های جدیدی برای شناسایی ریز RNA ها بر اساس روش های مشابه مورد استفاده در پژوهش حاضر انجام شده است. زنگ و همکاران (۲۰۲۰) از شبکه های عصبی پیچشی (CNN) و شبکه های عصبی مکرر (RNN) برای پیش بینی ریز RNA های انسانی استفاده کردند. برای هر دو مدل مقادیر حساسیت، اختصاصیت، معیار F و دقت بیشتر در محدوده ۸۰ تا ۹۰ درصد بود، در حالی که ضریب همبستگی متیو بین ۷۰ تا ۸۰ درصد برآورد شد (۵۳). کوردرو و همکاران (۲۰۱۹) با استفاده از شبکه های عصبی پیچشی براساس پردازش تصویر ریز RNA های پیش ساز، توالی های کد گذاری شده، ساختار توالی ها و بعضی از ویژگی های ترمودینامیکی در گونه های مختلف، مقدار صحت را ۹۲ درصد برآورد کردند (۹). فو و همکاران (۲۰۱۹) از ماشین بردار پشتیبان برای شناسایی ریز RNA ها از طریق اطلاعات دو جانبه توالی های ریز RNA های پیش ساز و ساختارهای ثانویه استفاده کردند. آنها با استفاده از الگوریتم ماشین بردار پشتیبان و تابع کرنل سیگموئید در داده های پایه، مقادیر حساسیت ۸۸ درصد، صحت ۹۰ درصد و ضریب همبستگی متیو را ۸۰ درصد برآورد کردند (۱۲). زنگ و همکاران (۲۰۱۹) از الگوریتم جنگل تصادفی که یک روش یادگیری ترکیبی برای دسته بندی می باشد، جهت شناسایی ریز RNA های مرتبط با بیماری ها استفاده کردند و میانگین های بدست آمده برای صحت و معیار F به ترتیب ۸۳ و ۸۲ درصد با استفاده از اعتبار سنجی متقاطع ۵ تایی برآورد شد (۵۵).

دانهدو و همکاران (۲۰۱۸) از روش یادگیری عمیق برای شناسایی اینکه توالی های RNA غیر کد کننده کوتاه، ریز RNA پیش ساز هستند یا اینکه به این گروه تعلق ندارند، استفاده کردند. آنها از ماتریس ساختار ثانویه پیش بینی شده به عنوان ورودی استفاده کردند و با معرفی آن به شبکه عصبی پیچشی دو بعدی، ویژگی های توالی ها را استخراج نمودند. نتایج آنالیز آنها بر روی داده های انسانی در ورودی های با سایز ثابت، مقدار حساسیت را ۸۷ درصد و معیار F را ۸۴ درصد نشان داد. در تحقیق آنها مقادیر حساسیت و معیار F برای ورودی های با سایز متغیر به ترتیب ۸۳ و ۸۵ درصد بدست آمد. همچنین آنها این آنالیز را برای سایر گونه ها نیز انجام دادند که در ورودی های با سایز ثابت این معیارها به ترتیب ۹۰ و ۸۹ درصد و برای ورودی هایی با سایز متغیر

۸۸ و ۸۹ درصد تخمین زده شد (۱۱).

به طور کلی ریز RNA ها حفاظت توالی بالایی در بین گونه ها نشان می دهند و از ساختار اولیه ساقه-حلقه در هسته تولید می شوند که از ویژگی های بسیار مهم ریز RNA ها می باشد. در روش های in-silico برای پیش بینی ریز RNA های جدید، انجام یک جستجوی همسانی در کل ژنوم کرم ابریشم لازم است. با توجه به ریز RNA های اولیه پیش فرض، این روش به طور کلی منجر به تعداد زیادی توالی می شود و سپس از طریق آنالیز ساختار ثانویه RNA، آنالیز نرم افزار پیش بینی و آنالیز دینامیک، غربال گری می شوند. تونگ و همکاران (۲۰۰۶) به جستجوی هومولوژی برای شناسایی ریز RNA های قبلاً تأیید شده پرداختند و ۲۴ ژن با پتانسیل ریز RNA را شناسایی کردند (۴۴). یو و همکاران (۲۰۰۸) ۱۱۴ ریز RNA حفاظت شده غیر تکراری را شناسایی کردند و ۱۴۸ ریز RNA از ژنوم کرم ابریشم را با یک سیستم پیش بینی پیشرفته بر اساس حلقه sRNA و ویژگی های ساختاری شناخته شده از ریز RNA های اولیه حیوانات شناسایی کردند (۵۲). یائو (۲۰۰۸)RNomics محاسباتی و آزمایشگاهی را برای پیش بینی و اعتبار یابی ریز RNA های کرم ابریشم بر اساس حفاظت توالی ریز RNA های بالغ و پیش سازهای آن ها با نشان دادن ساختارهای سنجاق سر به کار برد که منجر به شناسایی ۶۲ ریز RNA محفاظت شده بالقوه شد (۵۲). کائو و همکاران (۲۰۰۸) ۴۱ ریز RNA محافظت شده را با استفاده از یک رویکرد جستجوی همسانی محاسباتی شناسایی کردند که بیشتر آن ها انتخاب و هویت آن ها به صورت آزمایشگاهی تأیید شد (۸). بر اساس حفاظت توالی های ریز RNA، با استفاده از یک جستجوی همسانی محاسباتی بر اساس یک آنالیز توالی پیمایشی ژنومی، هوانگ و همکاران (۲۰۱۰)، ۱۶ ریز RNA جدید را شناسایی و توصیف کردند (۱۴). علاوه بر این هی و همکاران (۲۰۰۸) از ترکیبی از یک روش محاسباتی مبتنی بر جستجوهای همسانی توالی استفاده نمودند و از روش های آزمایشگاهی مبتنی بر ارزیابی ریزآرای و نوردرن بلات برای شناسایی آن ها استفاده کردند. با استفاده از این رویکرد ترکیبی، ۴۶ ریز RNA، ۲۱ ریز RNA محتمل و یک RNA کوچک جدید درکرم ابریشم مشخص شد. در میان نمونه های شناسایی شده، ۱۲ جفت ریز RNA[1] نیز وجود داشت (۱۳). در این پژوهش با استفاده از الگوریتم های مختلف بهترین مدل شناسایی ریز RNA های کرم ابریشم ارائه شد که تاکنون روی این گونه به طور اختصاصی این مطالعات انجام نشده است و با توجه به دقت بالای نتایج به دست آمده، استفاده از این مدل به عنوان یک روش با دقت بالا در صحت شناسایی ریز RNA های ژنوم کرم ابریشم پیشنهاد می شود.

نتیجه گیری کلی

به عنوان فاکتورهای کلیدی در تنظیم ژنی پس از رونویسی، ریز RNA ها به خوبی شناخته شده اند، زیرا عملکردهای تنظیمی بسیار مهم در فرآیندهای بیولوژیکی قابل توجهی مانند تمایز سلولی، تکثیر، آپوپتوز، رشد و بروز بیماری دارند.

تحقیقات کاربردی که به طور مستقیم در ریز RNA های کرم ابریشم انجام شده است، نشان داده است که ریز RNA ها می توانند اثرات قابل توجهی در مکانیسم های اساسی فرآیندهای رشد کرم های ابریشم ایجاد کنند. علاوه بر این پژوهش هایی که تاکنون انجام شده است، پایه پیشرفت در بهبود درک ما از شبکه های تنظیمی RNA و مکانیسم های مولکولی درگیر در الگوهای بیان ژن در طول مراحل مختلف زندگی را رائه می دهد. با توجه به تحقیقات محاسباتی ناکافی صورت گرفته در زمینه ریز RNA های کرم ابریشم، تحقیقات بیشتر در زمینه ریز RNA های این گونه، نشان دهنده پیشرفت مهمی در مطالعه RNA های غیر کدکننده دارای عملکرد های مهم بیولوژیکی در این گونه است که می تواند اطلاعات بیشتری در مورد فعالیت های RNA های غیر کد کننده ناشناخته ارائه دهد.

Refrences

1. Abbasi, V., M. R. Nasiri, and A. Javadmanesh. 2018. Prediction and In Silico Validation of Micro-RNAs in Different Tissues Originated from Ovine Chromosome 20. Iranian Journal of Animal Science Research, 11(2): 233-245. (In Persian).
2. Agarwal, S., C. Vaz, A. Bhattacharya, and A. Srinivasan. 2010. Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). BMC Bioinformatics, 11(1): S29.
3. Arowolo, M. O., M. Adebiyi, A. Adebiyi, and O. Okesola. 2020. PCA Model For RNA-Seq Malaria Vector Data Classification Using KNN And Decision Tree Algorithm. International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS), 1–8.
4. Bar, M., S. K. Wyman, B. R. Fritz, J. Qi, K. S. Garg, R. K. Parkin, E. M. Kroh, A. Bendoraite, P. S. Mitchell, and A. M. Nelson. 2008. MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. Stem Cells, 26(10): 2496–2505.
5. Ben-Hur, A., and J. Weston. 2010. A user's guide to support vector machines. In Data mining techniques for the life sciences Springer, Chapter 13, pages 223–239.

1- miRNA*s miRNA

6. Bentwich, I., A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, E. Sharon, Y. Spector, and Z. Bentwich. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. Nature Genetics, 37(7): 766–770.
7. Bhaskar, H., D. C. Hoyle, and S. Singh. 2006. Machine learning in bioinformatics: A brief survey and recommendations for practitioners. Computers in Biology and Medicine, 36(10): 1104–1125.
8. Cao, J., C. Tong, X. Wu, J. Lv, Z. Yang, and Y. Jin. 2008. Identification of conserved microRNAs in Bombyx mori (silkworm) and regulation of fibroin L chain production by microRNAs in heterologous system. Insect Biochemistry and Molecular Biology, 38(12): 1066–1071.
9. Cordero, J., V. Menkovski, and J. Allmer. 2019. Detection of pre-microRNA with Convolutional Neural Networks. bioRxiv, Europe PMC, 1-12.
10. Ding, J., S. Zhou, and J. Guan. 2010. MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. BMC Bioinformatics, 11 Suppl 1(Suppl 11): S11.
11. Do, B. T., V. Golkov, G. E. Gürel, and D. Cremers. 2018. Precursor microRNA Identification Using Deep Convolutional Neural Networks. BioRxiv, 414656.
12. Fu, X., W. Zhu, L. Cai, B. Liao, L. Peng, Y.Chen, and J. Yang. 2019. Improved pre-miRNAs identification through mutual information of pre-miRNA sequences and structures. Frontiers in Genetics, 10(FEB): 1–12.
13. He, P., Z. Nie, J. Chen, Z. Lv, Q. Sheng, S. Zhou, X. Gao, L. Kong, and X. Wu. 2008. Identification and characteristics of microRNAs from Bombyx mori. BMC Genomics, 9(1): 248.
14. Huang, Y., Q. Zou, S. Tang, L. Wang, and X. Shen. 2010. Computational identification and characteristics of novel microRNAs from the silkworm (Bombyx mori L.). Molecular Biology Reports. 37: 3171–3176.
15. Jabbar, M. A., and S. Samreen. 2016. Heart disease prediction system based on hidden naïve bayes classifier. International Conference on Circuits, Controls, Communications and Computing (I4C): 1–5.
16. Jiang, P., H. Wu, W. Wang, W.Ma, X. Sun, and Z. Lu. 2007. MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Research, 35(SUPPL.2): W339-W344.
17. Kadri, S., V. Hinman, and P. V. Benos. 2009. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. BMC Bioinformatics, 10(Suppl 1): S35.
18. Kozomara, A., M. Birgaoanu, and S. Griffiths-Jones. 2018. miRBase: from microRNA sequences to function. Nucleic Acids Research, 47(D1): D155–D162.
19. Lai, E. C., P. Tomancak, R. W. Williams, and G. M. Rubin. 2003. Computational identification of Drosophila microRNA genes. Genome Biology, 4(7): R42.
20. Larranaga, P., B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafé, and A. Pérez. 2006. Machine learning in bioinformatics. Briefings in Bioinformatics, 7(1): 86–112.
21. Li, L., J. Xu, D. Yang, X. Tan, and H. Wang. 2010. Computational approaches for microRNA studies: a review. Mammalian Genome, 21(1–2): 1–12.
22. Li, S.C., C.K. Shiau, and W. Lin. 2007. Vir-Mir db: prediction of viral microRNA candidate hairpins. Nucleic Acids Research, 36(suppl_1): D184–D189.
23. Lim Lee, P., C. Lau Nelson, G. Weinstein Earl, Y. S. Abdelhakim Aliaa, W. Rhoades Matthew, B. Burge Christopher, and P. Bartel David. 2003. The microRNAs of Caenorhabditis elegans. Genes Dev, 17(8): 991–1008.
24. Lindow, M., and J. Gorodkin. 2007. Principles and limitations of computational microRNA gene and target finding. DNA and Cell Biology, 26(5): 339–351.
25. Liu, C.G., G. A. Calin, B.Meloon, N. Gamliel, C. Sevignani, M. Ferracin, C. D. Dumitru, M. Shimizu, S. Zupo, and M. Dono. 2004. An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. Proceedings of the National Academy of Sciences, 101(26): 9740–9744.
26. Lou, S., T. Sun, H. Li, and Z. Hu. 2018. Mechanisms of microRNA-mediated gene regulation in unicellular model alga Chlamydomonas reinhardtii. Biotechnology for Biofuels, 11(1): 244.
27. Magyar, L. 2018. A Review of the Utility of Bayesian Network Models. The University of Akron, ideaexchange.uakron.edu.1-28.
28. Mendes, N. D., A. T. Freitas, and M.F. Sagot. 2009. Current tools for the identification of miRNA genes and their targets. Nucleic Acids Research, 37(8): 2419–2433.
29. Milagro, F. I., J. Miranda, M. P. Portillo, A. Fernandez-Quintela, J. Campion, and J. A. Martínez. 2013. High-throughput sequencing of microRNAs in peripheral blood mononuclear cells: identification of potential weight loss biomarkers. PloS One, 8(1): e54319.
30. Nam, J. W., J. Kim, S. K. Kim, and B. T. Zhang. 2006. ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. Nucleic Acids Research, 34(suppl_2): W455–W458.
31. Nam, J. W., K. R. Shin, J. Han, Y. Lee, V. N. Kim, and B. T. Zhang. 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Research, 33(11): 3570–3581.
32. Nelson, P. T., D. O. N. A. Baldwin, W. P. Kloosterman, S. Kauppinen, R. H. A. Plasterk, and Z. Mourelatos .2006.

RAKE and LNA-ISH reveal microRNA expression and localization in archival human brain. Rna, 12(2): 187–191.
33. Ng, K. L. S., and S. K. Mishra. 2007. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics, 23(11): 1321–1330.
34. Ntranos, V., L. Yi, P. Melsted, and L. Pachter. 2019. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. Nature Methods, 16(2): 163-166.
35. Oulas, A., A. Boutla, K. Gkirtzou, M. Reczko, K. Kalantidis, and P. Poirazi. 2009. Prediction of novel microRNA genes in cancer-associated genomic regions—a combined computational and experimental approach. Nucleic Acids Research, 37(10): 3276–3287.
36. Paicu, C., I. Mohorianu, M. Stocks, P. Xu, A. Coince, M. Billmeier, T. Dalmay, V. Moulton, and S. Moxon. 2017. miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing datasets. Bioinformatics, 33(16): 2446-2454.
37. Ritchie, W., D. Gao, and J. E. J. Rasko. 2012. Defining and providing robust controls for microRNA prediction. Bioinformatics, 28(8): 1058–1061.
38. Saçar, M. D., and J. Allmer. 2014. Machine learning methods for microRNA gene prediction. In miRNomics: MicroRNA Biology and Computational Analysis. Springer, 1107:177-87
39. Saçar, M. D., H. Hamzeiy, and J. Allmer. 2013. Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins? Journal of Integrative Bioinformatics, 10(2): 1–11.
40. Sheng, Y., P. G. Engström, and B. Lenhard. 2007. Mammalian microRNA prediction through a support vector machine model of sequence and structure. PloS One, 2(9): e946.
41. Singh, S., and R. Singh. 2017. Application of supervised machine learning algorithms for the classification of regulatory RNA riboswitches. Briefings in Functional Genomics, 16(2): 99–105.
42. Siomi, H., and M. C. Siomi. 2010. Posttranscriptional regulation of microRNA biogenesis in animals. Molecular Cell, 38(3): 323–332.
43. Terai, G., T. Komori, K. Asai, and T. Kin. 2007. miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity. Rna, 13(12): 2081–2090.
44. Tong, C., Y. Jin, and Y. Zhang. 2006. Computational prediction of microRNA genes in silkworm genome. Journal of Zhejiang University Science B, 7(10): 806–816.
45. Tran, V. D. T., S. Tempel, B. Zerath, F. Zehraoui, and F. Tahi. 2015. miRBoost: boosting support vector machines for microRNA precursor classification. RNA (New York, N.Y.), 21(5): 775-785.
46. Várallyay, E., J. Burgyán, a nd Z. Havelda. 2007. Detection of microRNAs by Northern blot analyses using LNA probes. Methods, 43(2): 140–145.
47. Wang, X., S. M. Tang, and X. J. Shen. 2014. Overview of research on Bombyx mori microRNA. Journal of Insect Science, 14(133): 133.
48. Wu, Y., B. Wei, H. Liu, T. Li, and S. Rayner. 2011. MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. BMC Bioinformatics, 12(1): 107.
49. Xue, C., F. Li, T. He, G.P. Liu, Y. Li, and X. Zhang. 2005. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics, 6: 310.
50. Xue, H., Z. Wei, K. Chen, Y. Tang, X. Wu, J. Su, and J. Meng. 2020. Prediction of RNA methylation status from gene expression data using classification and regression methods. Evolutionary Bioinformatics, 16: 1176934320915707.
51. Yousef, M., S. Jung, A. V. Kossenkov, L. C. Showe, and M. K. Showe. 2007. Naïve Bayes for microRNA target predictions—machine learning for microRNA targets. Bioinformatics, 23(22): 2987–2992.
52. Yu, X., Q. Zhou, S.C. Li, Q. Luo, Y. Cai, W. Lin, H. Chen, Y. Yang, S. Hu, and J.Yu. 2008. The silkworm (Bombyx mori) microRNAs and their expressions in multiple developmental stages. PloS One, 3(8): e2997.
53. Zhang, G., Y. Deng, Q. Liu, B. Ye, Z. Dai, Y. Chen, and X. Dai. 2020. Identifying circular RNA and predicting its regulatory interactions by machine learning. Frontiers in Genetics, 11: 655.
54. Zhang, Y. Q., J. C. Rajapakse, and B. T. Zhang. 2008. Supervised Learning Methods for MicroRNA Studies. Machine Learning in Bioinformatics, Chapter 16, page 339.
55. Zheng, K., Z. H. You, L. Wang, Y. Zhou, L. P. Li, and Z. W. Li. 2019. MLMDA: A machine learning approach to predict and validate MicroRNA-disease associations by integrating of heterogenous information sources. Journal of Translational Medicine, 17(1): 1–14.
56. Zheng, X., X. Fu, K. Wang, and M. Wang. 2020. Deep neural networks for human microRNA precursor detection. BMC Bioinformatics, 21(1): 1-7.
57. Zhong, L., and J. T. L. Wang. 2016. Effective Classification of MicroRNA Precursors Using Combinatorial Feature Mining and AdaBoost Algorithms. ArXiv:1610.02281,ui.adsabs.harvard.edu.

Iranian Journal of Animal Science Research
Vol. 13, No. 4, Winter 2022, p. 615-627

نشریه پژوهشهای علوم دامی ایران
جلد ۱۳، شماره ۴، زمستان ۱۴۰۰، ص. ۶۲۷-۶۱۵

A survey on effect of multilayer perceptron on the accuracy of selection of silkworm (*Bombyx mori*) microRNA genes

Atefeh Seyeddokht[1*], Javad Rahmaninia[2]
Submitted: 07-03-2020
Accepted: 19-12-2020

Seyeddokht, A., and J. Rahmaninia. 2022. A survey on effect of multilayer perceptron on the accuracy of selection of silkworm (*Bombyx mori*) microRNA genes. Iranian Journal of Animal Science Research 13(4):615-627.

Introduction MicroRNAs (miRNAs) constitute a large family of non-protein-coding small RNA (ncRNA) molecules and have important roles in the regulation of both plant and animal developmental procedures. Generally, sequences of miRNA demonstrate high sequence conservation across animals and are produced from the primary stem-loop structure in the nucleus, which is an important feature of miRNAs. MiRNAs are one of the most important regulatory factors involved in post-transcriptional levels of gene expression that contribute to the modulation of a large number of physiological processes such as development, metabolism and disease occurrence. To date, A few studies related to miRNAs of the economically important silkworm, *Bombyx mori*, have been carried out, focusing on detection, expression study, and prediction of function. Machine learning approaches are crucial for prediction success. These methods can solve classification problem.

Materials and Method Although hundreds of miRNAs have been detected in different animals, a lot of them are still unknown. Then, finding of novel miRNA genes is an essential step for understanding miRNA intervened post transcriptional regulation processes. It appears that biological methods to recognize miRNA genes might be inadequate in their capacity to identify uncommon miRNAs and are further limited to the tissues surveyed and the developmental phase of the animal under experiment. These restrictions have led to the development of new computational methods attempting to detect potential miRNAs. Experimentally verified miRNA sequences in miRBase release 22.0 were extracted for inclusion in the positive data set. In the miRBase, the reported secondary structures were predicted by a collection of RNA folding software packages. Consequently, in this study for uniformity, all miRNA secondary structures analyzed using RNAfold packages. The major step for machine learning approaches is the selection of a suitable negative dataset. It is important for a well-trained classifier. If the sequences are too artificial, e.g. completely random sequences, then there is a risk that the classifiers will not be well trained to differentiate between different categories of real biological sequences. Conversely, if the negative dataset is too similar to the positive dataset, the classifiers will be unable to find a way to adequately differentiate between these two data sets. We investigated several different types of negative sequences and finally selected negative sequences which made the best distinction with positive data set. The positive training dataset for our classifier development composed of known silkworm pre miRNAs, while the negative training dataset composed of other ncRNA sequences. Our feature set composed of various features and selecting the most discriminative set of features would increase the performance, efficiency and comprehensibility of a classifier method by reducing its complexity.

Results and Discussion Secondary structural patterns of pre miRNA used in this study such as the intramolecular base pairing of pre miRNA is an important beneficial feature for miRNAs classification. The selective powers of the two different classes of miRNAs secondary structural conformation (dot-bracket notation) were analyzed. Secondary structural feature of miRNA such as Minimum Free Energy, Watson-crick base pairing (AU, GC), Wobble base pairing (G-U) and unpaired bases (A, G, C, U) is analyzed by different algorithms. Here

1-Animal Science Research Department, Khorasan Razavi Agricultural and Natural Resources Research and Education Center, AREEO, Mashhad, Iran
2-Animal Science Research Institute of Iran, Agricultural Research, Education and Extension Organization (AREEO), Karaj, Iran
*Corresponding Author: Email: a.seyeddokht@areeo.ac.ir
Doi:10.22067/ijasr.2020.38276.0

we could successfully solve classification problem by developing an effective classification system using machine learning techniques. Our approach includes introducing more representative datasets, extracting new effective biological features, and comprehensive evaluating of classification performance through these methods via cross-validation. Performance of different algorithms was measured by the total number of true negatives (TN), true positives (TP), false positives (FP), false negatives (FN), and accuracy (ACC). In order to evaluate the efficiency of various methods developed in this study, various parameters like F-measure, Matthews correlation coefficient (MCC), accuracy (ACC) and, ROC area were calculated. Performance measurement of various models tested with data from miRBase in release 22 in ten-fold cross validation. Multilayer Perceptron model could predict pre miRNAs from non-coding sequences that can be important for detecting the true pre miRNAs in genomic sequences. Consequently a new method on miRNA prediction model could be favorable to understand the characteristics miRNA associated with miRNA biogenesis.

Conclusion Research on miRNA represents important progress in the study of ncRNAs and may provide further information on understanding of RNA regulation networks. Practical research on silkworm microRNAs has shown that microRNAs can have significant effects on the underlying mechanisms of silkworm growth processes. In addition to the research that has been done so far, it provides the basis for advances in improving our understanding of RNA regulatory networks and the molecular mechanisms involved in gene expression patterns during different stages of silkworm life. Due to insufficient computational research in the field of silkworm microRNAs, further research on the microRNAs of this species represents an important advance in the study of noncoding RNAs, which can provide further information on the activity of noncoding RNAs. Machine learning algorithms will help the researcher discover the uncover miRNA that many researchers were not able to explore.

Key words: Computational Methods, MicroRNA, Regulatory Factors, Silkworm.